

# A Bimodal Co-Sparse Analysis Model for Image Processing

Martin Kiechle      Tim Habigt      Simon Hawe  
Martin Kleinstein

Department of Electrical Engineering and Information Technology,  
Technische Universität München, Munich, Germany

{martin.kiechle,tim,simon.hawe,kleinstein}@tum.de

<http://www.gol.ei.tum.de>

## Abstract

*The success of many computer vision tasks lies in the ability to exploit the interdependency between different image modalities such as intensity and depth. Fusing corresponding information can be achieved on several levels, and one promising approach is the integration at a low level. Moreover, sparse signal models have successfully been used in many vision applications. Within this area of research, the so-called co-sparse analysis model has attracted considerably less attention than its well-known counterpart, the sparse synthesis model, although it has been proven to be very useful in various image processing applications.*

*In this paper, we propose a co-sparse analysis model that is able to capture the interdependency of two image modalities. It is based on the assumption that a pair of analysis operators exists, so that the co-supports of the corresponding bimodal image structures are correlated. We propose an algorithm that is able to learn such a coupled pair of operators from registered and noise-free training data. Furthermore, we explain how this model can be applied to solve linear inverse problems in image processing and how it can be used for image registration tasks. This paper extends the work of some of the authors by two major contributions. Firstly, a modification of the learning process is proposed that a priori guarantees unit norm and zero-mean of the rows of the operator. This accounts for the intuition that contrast in image modalities carries the most information. Secondly, the model is used in a novel bimodal image registration algorithm which estimates the transformation parameters of unregistered images of different modalities.*

## 1 Introduction

In the past, the majority of methods tackling problems in computer vision were focused on working on a single image modality, typically a color or grayscale image captured with a digital camera. Due to the progress in sensor technologies, sensors that capture different types of image modalities beyond intensity, have become affordable and popular. Well-known examples of multi-modal image sensors include thermal,

multispectral and depth cameras, as well as MRI or PET. These image signals often carry information about one another and exploiting this interdependency is beneficial for solving problems in computer vision, such as reconstruction, registration, segmentation, detection, or recognition in a more robust way. Inspired by biological systems which perceive their environment through many different signal modalities at once, fusing sensory information from different modalities has emerged as an important research topic. Existing fusion schemes can be grouped according to their level of fusion. Methods of decision-level fusion work independently on the different modalities to make separate task-dependent decisions, which are then fused according to a certain rule or confidence measure. Feature-level fusion methods integrate modality-specific features to derive a decision, for instance the well-known bag-of-words method in object classification. The method presented in this paper, belongs to the group of low-level fusion, where the multimodal information is integrated on the pixel level.

Often, low-level integration is interpreted as finding a mapping from one modality or image domain to another. Typically, this mapping is learned from sets of aligned local image patches to make corresponding algorithms computationally tractable [12, 2, 14, 20]. More recent approaches aim at capturing the low-level integration across modalities via sparse coding, where the interdependencies of the signals are reflected in interdependencies of their sparse codes. This concept is used in several methods to find a mapping between different resolution levels or across image modalities. In [34], Yang *et al.* apply such a scheme to single image super-resolution (SR). They learn two dictionaries for corresponding low-resolution (LR) and high-resolution (HR) image patches and fuse the two domains through a common sparse representation. In [19], Li *et al.* propose a SR approach across the two different image modalities intensity and depth. Three domains are fused through separate dictionaries for LR and HR depth as well as HR intensity patches. The dictionaries are learned by enforcing common support in the sparse representation.

The assumption of a common sparse code across the different domains is often too strict in practice. In [22], a less restrictive model is proposed, in which a dictionary is learned for the source image domain together with a transformation matrix which transforms the sparse representations of the source domain to signals in the target domain. Wang *et al.* [32] use linear regression between the sparse representations over different dictionaries for the image domains. A similar idea is followed by Jia *et al.* [15], who refine the linear mapping of sparse codes by a local parameter regression for different subsets of sparse representations. While all of these fusion methods rely on the sparse synthesis model, the related co-sparse analysis model [10] has not been considered yet in such a multi-modal setting. This is particularly surprising given its excellent performance in unimodal image processing tasks [24, 13, 33].

In this work, we propose a bimodal data model based on co-sparsity for two image modalities. It is based on an extension of the co-sparse analysis model and allows to find signal representations that are simultaneously co-sparse across the two different image domains. We revisit the learning procedure proposed by some of the authors in [16] with refinements on parameter selection and a modification of the manifold structure on which the model is learned. In order to demonstrate both, descriptive power and cross-modal coupling of this model, we first propose to employ it as a prior for solving inverse problems, which we subsequently validate in an image guided depth-map reconstruction task. The model is further applied in a novel algorithm for bimodal image registration, which, to the best of our knowledge, is the first sparsity-based approach to tackle this problem. Therein, we combine the proposed joint bimodal co-sparsity model with an optimization on Lie groups and achieve favorable results in comparison to other image registration methods.

We outline the remainder of this paper as follows. In Section 2, the bimodal co-sparse analysis model is described and an appropriate learning objective is derived. Subsequently, we explain in Section 3 how a solution of this learning objective can be computed efficiently using optimization techniques on matrix manifolds. Sections 4 and Section 5 contain the experiments on bimodal image reconstruction and bimodal

image registration.

## 2 Bimodal Co-Sparse Analysis Model

The (unimodal) co-sparse analysis model [24] assumes that for a given class of signals  $\mathcal{S} \subset \mathbb{R}^n$ , there exists a so-called *analysis operator*  $\mathbf{\Omega} \in \mathbb{R}^{k \times n}$  such that the *analyzed vector*

$$\mathbf{\Omega}\mathbf{s} \text{ is sparse for all } \mathbf{s} \in \mathcal{S}. \quad (1)$$

From a geometrical perspective,  $\mathcal{S}$  is contained in a union of subspaces and  $\mathbf{s} \in \mathcal{S}$  lies in the intersection of all hyperplanes whose normal vectors are given by the rows of  $\mathbf{\Omega}$  that are indexed by the zero entries of  $\mathbf{\Omega}\mathbf{s}$ . This index set is called the *co-support* of  $\mathbf{s}$  and is denoted by

$$\text{cosupp}(\mathbf{\Omega}\mathbf{s}) := \{j \mid (\mathbf{\Omega}\mathbf{s})_j = 0\}, \quad (2)$$

where  $(\mathbf{\Omega}\mathbf{s})_j$  is the  $j$ -th entry of the analyzed vector. In image processing applications,  $\mathcal{S}$  typically consists of vectorized image patches. One prominent example for a co-sparse analysis model for natural images is the so-called total variation operator. This ad-hoc model assumes that differences of neighboring pixel intensities result in a sparse vector. However, it has been shown that such ad-hoc models are inferior to models that are adapted to the specific class  $\mathcal{S}$  of interest, cf. [13, 28, 33]. Consequently, *analysis operator learning* aims at finding the most suitable analysis operator for a given class  $\mathcal{S}$ .

In this work, we consider two signal classes  $\mathcal{S}_U$  and  $\mathcal{S}_V$  of different modalities that emanate from the same physical object. Consider for example an intensity image and a depth map captured from the same scene. More precisely, let  $(\mathbf{s}_U, \mathbf{s}_V) \in \mathcal{S}_U \times \mathcal{S}_V$ . We assume that these signal pairs  $(\mathbf{s}_U, \mathbf{s}_V)$  allow a co-sparse representation with an appropriate pair of analysis operators  $(\mathbf{\Omega}_U, \mathbf{\Omega}_V) \in \mathbb{R}^{k \times n_U} \times \mathbb{R}^{k \times n_V}$ . Based on the knowledge that the structure of a signal is encoded in its co-support (2), we assume that *a pair of analysis operators exists such that the co-supports of  $\mathcal{S}_U$  and  $\mathcal{S}_V$  are statistically dependent*. The bimodal co-sparse analysis model is thus based on the assumption that the conditional probability of  $j$  belonging to the co-support of  $\mathbf{s}_V$  given that  $j$  belongs to the co-support of  $\mathbf{s}_U$  is significantly higher than the unconditional probability, i.e.

$$\Pr(\{j \in \text{cosupp}(\mathbf{\Omega}_V \mathbf{s}_V)\} \mid \{j \in \text{cosupp}(\mathbf{\Omega}_U \mathbf{s}_U)\}) \gg \Pr(\{j \in \text{cosupp}(\mathbf{\Omega}_V \mathbf{s}_V)\}). \quad (3)$$

Geometrically interpreted, we aim at partitioning the signal space for each of the two modalities in such a way, that the partitions not only represent subsets of signals of interest but simultaneously relate to a partition of the other modality.

Clearly, this model is idealized, since in practice the entries of the analyzed vectors are not exactly equal to zero. In the following, we relax this strict co-sparsity assumption and show how a coupled pair of analysis operators can be jointly learned, such that aligned bimodal signals analyzed by these operators adhere to the co-sparse model.

More specifically, we aim at learning the coupled pair of bimodal analysis operators  $(\mathbf{\Omega}_U, \mathbf{\Omega}_V) \in \mathbb{R}^{k \times n_U} \times \mathbb{R}^{k \times n_V}$  for two signal modalities. Therefore, we use a set of  $M$  aligned and corresponding training pairs

$$\{(\mathbf{s}_U^{(i)}, \mathbf{s}_V^{(i)}) \in \mathbb{R}^{n_U} \times \mathbb{R}^{n_V}\}_{i=1}^M. \quad (4)$$

For simplicity, we assume throughout this work that training signals of both modalities have the same size, i.e.  $n_U = n_V = n$ . Now, we incorporate the proposed condition (3) into the learning process by inducing

the zeros of corresponding analyzed vectors  $\Omega_U \mathbf{s}_U^{(i)}, \Omega_V \mathbf{s}_V^{(i)}$  to be at the same positions. From here on, the function

$$\mathbf{x} \mapsto \sum_{j=1}^k \log(1 + \nu x_j^2), \quad (5)$$

with  $\nu > 0$  being a positive weight and  $x_j$  denoting the entries of  $\mathbf{x}$ , serves as an appropriate sparsity measure. Note, that any other smooth sparsity measure principally leads to similar results. With this, the coupled sparsity is controlled through the function

$$g(\Omega_U \mathbf{s}_U^{(i)}, \Omega_V \mathbf{s}_V^{(i)}) := \sum_{j=1}^k \log \left( 1 + \nu ((\Omega_U \mathbf{s}_U^{(i)})_j^2 + (\Omega_V \mathbf{s}_V^{(i)})_j^2) \right). \quad (6)$$

Multi-modal dictionary learning methods which model the coupling across modalities by the same sparse representation have found to be too restrictive or require dictionaries of high over-completeness [15]. Although we also attain the coupling over the sparse representation, there are two key differences. First, the co-sparse analysis model tends to lead to a richer union of subspaces than the synthesis model [24] and is therefore inherently less restrictive. Second, we relax the strict coupling of a common sparse representation by the smooth function (6), which less restrictively promotes correlated representations. As it turns out, this results in substantial cross-modal coupling without the need of highly redundant operators.

To find the ideal pair of bimodal operators we minimize the empirical expectation of (6) over all training signal pairs, which reads as

$$G(\Omega_U, \Omega_V) := \frac{1}{M} \sum_{i=1}^M g(\Omega_U \mathbf{s}_U^{(i)}, \Omega_V \mathbf{s}_V^{(i)}). \quad (7)$$

In order to avoid the trivial solution, some regularization constraints on  $\Omega$  have to be imposed. In a first step, we demand the rows of  $\Omega$  to have unit Euclidean norm, i.e. we restrict the transpose of possible solutions to the so-called *oblique manifold*

$$\Omega^\top \in \text{OB}(n, k) := \mathbb{S}_{n-1}^{\times k}, \quad (8)$$

where  $\mathbb{S}_{n-1}$  denotes the unit sphere in  $\mathbb{R}^n$ . Unfortunately, in general such an operator  $\Omega$  is biased by signals with varying mean values, i.e.

$$\Omega \mathbf{s} \neq \Omega(\mathbf{s} + c \mathbb{1}_n), \quad (9)$$

where  $\mathbb{1}_n$  is the vector with all entries being equal to 1. Because we are interested in the structure of an image signal independent of its constant component, we need to avoid the influence of different mean values. For natural images, this can for instance be interpreted as a certain invariance to changes in brightness. A straightforward and popular way to account for such illumination invariance is to learn the model from zero-mean training signals  $\mathbf{s}_c = \mathbf{s} - \bar{\mathbf{s}}$  by subtracting the mean. If we denote  $\mathbf{I}_n$  as the identity operator and  $\mathbf{J}_n$  as a matrix with all elements equal to one, we can express this centering operation in matrix vector notation as

$$\mathbf{s}_c = (\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n) \mathbf{s}. \quad (10)$$

This operation, however, induces a new trivial solution  $\frac{1}{\sqrt{n}} \mathbb{1}_n$  for the rows  $\omega^\top$  of the operator without encoding any structural information. Therefore, we extend our approach in [16] by further restricting the

rows of possible solutions to the orthogonal complement of  $\mathbb{1}_n$ , which we shortly denote by  $\mathbb{1}_n^\perp$ . Thus, we further restrict the transpose of admissible solutions  $\Omega$  to

$$\mathcal{R} = (\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp)^{\times k}. \quad (11)$$

Besides these constraints, it has been well investigated [13] that coherence and rank are important properties of an analysis operator to well represent a signal class. Therefore we regularize the rank of  $\Omega|_{\mathbb{1}_n^\perp}$  with the penalty function

$$h(\Omega) := -\frac{1}{(n-1)\log(n-1)} \log \det\left(\frac{1}{k} \mathbf{W}^\top \Omega^\top \Omega \mathbf{W}\right), \quad (12)$$

in which the columns of  $\mathbf{W} \in \mathbb{R}^{n \times (n-1)}$  form an arbitrary orthonormal basis of  $\mathbb{1}_n^\perp$ .

Furthermore, to control the coherence of the operator and enforce distinct rows, we adopt the regularizer proposed in [13], which is a log-barrier function on the scalar product of all operator rows, i.e.

$$r(\Omega) := -\sum_{1 \leq i < l \leq k} \log(1 - (\omega_i^\top \omega_l)^2). \quad (13)$$

The combination of the two regularizers

$$p(\Omega) := \kappa h(\Omega) + \mu r(\Omega), \quad (14)$$

with  $\kappa, \mu \in \mathbb{R}^+$  being positive weights, in conjunction with the co-sparsity objective function (7) comprises our final learning function

$$L(\Omega_U, \Omega_V) := G(\Omega_U, \Omega_V) + p(\Omega_U) + p(\Omega_V). \quad (15)$$

Accordingly, the problem of finding the appropriate pair of joint bimodal co-sparse analysis operators is stated as

$$(\Omega_U^\top, \Omega_V^\top) \in \arg \min_{\mathbf{X}_U, \mathbf{X}_V \in \mathcal{R}} L(\mathbf{X}_U^\top, \mathbf{X}_V^\top). \quad (16)$$

### 3 Joint Bimodal Analysis Operator Learning Algorithm

In order to find an optimal solution to (16), we employ a conjugate gradient method on manifolds. To make this work self-contained, we briefly review conjugate gradient and gradient descent methods on matrix manifolds in general and point the interested reader to [1] for further details. Based on this concept, we then derive the learning algorithm for the proposed joint bimodal analysis operators in Section 3.2. The proposed optimization framework will also prove useful for the reconstruction and alignment algorithms presented in Section 4 and Section 5, respectively.

#### 3.1 Line Search Methods on Matrix Manifolds

Let  $\mathcal{M}$  be a smooth Riemannian sub-manifold of a finite dimensional real vector space  $\mathbb{V}$  with a scalar product  $\langle \cdot, \cdot \rangle$  and consider the problem of minimizing a smooth real valued function

$$f: \mathcal{M} \rightarrow \mathbb{R}. \quad (17)$$

The general idea of line search methods like conjugate gradient or gradient descent algorithms on manifolds is that, starting from some point  $\mathbf{X} \in \mathcal{M}$  we search along a curve on the manifold towards the minimizer

of (17). In our setting, the descent direction is an element of the tangent space  $T_{\mathbf{X}}\mathcal{M}$ , and we search for the updated iterate along geodesics. In the case where  $f$  is defined in the embedding space  $\mathbb{V}$ , its gradient  $\nabla f(\mathbf{X})$  with respect to  $\langle \cdot, \cdot \rangle$  is uniquely determined by

$$\left. \frac{d}{dt} \right|_{t=0} f(\mathbf{X} + t\mathbf{H}) = \langle \nabla f(\mathbf{X}), \mathbf{H} \rangle \quad \text{for all } \mathbf{H} \in \mathbb{V}. \quad (18)$$

The *Riemannian gradient*  $\mathbf{G}(\mathbf{X})$ , which serves as the (negative) search direction for a gradient descent method on manifolds, is simply the orthogonal projection of  $\nabla f(\mathbf{X})$  onto the tangent space  $T_{\mathbf{X}}\mathcal{M}$ , i.e.

$$\mathbf{G}(\mathbf{X}) = \Pi_{T_{\mathbf{X}}\mathcal{M}}(\nabla f(\mathbf{X})), \quad (19)$$

with  $\Pi_{T_{\mathbf{X}}\mathcal{M}}$  denoting the orthogonal projection with respect to  $\langle \cdot, \cdot \rangle$ . Now let  $t \mapsto \Gamma(\mathbf{X}, \mathbf{H}, t)$  denote the geodesic emanating from  $\mathbf{X} \in \mathcal{M}$  in direction  $\mathbf{H} \in T_{\mathbf{X}}\mathcal{M}$ , that is

$$\Gamma(\mathbf{X}, \mathbf{H}, 0) = \mathbf{X} \quad \text{and} \quad \left. \frac{d}{dt} \right|_{t=0} \Gamma(\mathbf{X}, \mathbf{H}, t) = \mathbf{H}. \quad (20)$$

Schematically, line search methods on manifolds update the  $i$ -th estimate  $\mathbf{X}^i$  by

$$\mathbf{X}^{i+1} = \Gamma(\mathbf{X}^i, \mathbf{H}^i, t^i), \quad (21)$$

where  $\mathbf{H}^i \in T_{\mathbf{X}^i}\mathcal{M}$  is the descent direction and  $t^i \in \mathbb{R}$  is a suitable step-size. If  $\mathbf{H}^i = -\mathbf{G}(\mathbf{X}^i)$  is the negative Riemannian gradient, the method is a gradient descent algorithm and provides linear convergence to a local minimizer for appropriate step-size selections [1].

In practice, faster convergence can often be achieved by adapting conjugate gradient methods to the manifold setting. In this case, the search direction  $\mathbf{H}^{i+1} \in T_{\mathbf{X}^{i+1}}\mathcal{M}$  is a linear combination of the Riemannian gradient  $\mathbf{G}^{i+1} := \mathbf{G}(\mathbf{X}^{i+1}) \in T_{\mathbf{X}^{i+1}}\mathcal{M}$  and the previous search direction  $\mathbf{H}^i$ . Since the linear combinations of elements from different tangent spaces are not defined, the parallel transport along geodesics is employed to identify the different tangent spaces. If we denote this parallel transport by

$$\Psi_i^{i+1}: T_{\mathbf{X}^i}\mathcal{M} \rightarrow T_{\mathbf{X}^{i+1}}\mathcal{M}, \quad (22)$$

the conjugate gradient method on manifold updates the search direction via

$$\mathbf{H}^{i+1} := -\mathbf{G}^{i+1} + \beta^i \Psi_i^{i+1}(\mathbf{H}^i), \quad (23)$$

where initially,  $\mathbf{H}^0 := -\mathbf{G}^0$ . For our purposes, the update parameter  $\beta^i$  is chosen according to a manifold adaption of the Fletcher-Reeves and Dai-Yuan formula. More precisely, we employ a hybridization of the Hestenes-Stiefel Formula and the Dai Yuan formula

$$\beta_{hyb}^{(i)} = \max(0, \min(\beta_{DY}^{(i)}, \beta_{HS}^{(i)})), \quad (24)$$

which has been suggested in [8], where

$$\beta_{HS}^{(i)} = \frac{\langle \mathbf{G}^{(i+1)}, \mathbf{G}^{i+1} - \Psi_i^{i+1}(\mathbf{G}^i) \rangle}{\langle \Psi_i^{i+1}(\mathbf{H}^i), \mathbf{G}^{i+1} - \Psi_i^{i+1}(\mathbf{G}^i) \rangle}, \quad (25)$$

$$\beta_{DY}^{(i)} = \frac{\langle \mathbf{G}^{(i+1)}, \mathbf{G}^{(i+1)} \rangle}{\langle \Psi_i^{i+1}(\mathbf{H}^i), \mathbf{G}^{i+1} - \Psi_i^{i+1}(\mathbf{G}^i) \rangle}. \quad (26)$$

### 3.2 Geometric Conjugate Gradient for Joint Bimodal Analysis Operator Learning

We propose a geometric conjugate gradient method as explained in the previous subsection to tackle problem (16) for learning the joint bimodal analysis operator. First, we have to ensure that  $\mathcal{R}$  as given in (11) is indeed a manifold.

*Lemma.* The set  $\mathcal{R} = (\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp)^{\times k}$  is a Riemannian submanifold of  $\mathbb{R}^{n \times k}$  and the tangent space at  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_k] \in \mathcal{R}$  is given by

$$T_{\mathbf{X}}\mathcal{R} = T_{\mathbf{x}_1}(\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp) \times \dots \times T_{\mathbf{x}_k}(\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp), \quad (27)$$

with

$$T_{\mathbf{x}}(\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp) = \{\mathbf{h} \in \mathbb{R}^n \mid \mathbf{h}^\top [\mathbf{x}, \mathbb{1}_n] = 0\}. \quad (28)$$

*Proof.* By using the product manifold structure, it is sufficient to show that  $\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp$  is a submanifold of  $\mathbb{R}^n$  with its tangent space as given in (28). Consider the function

$$F: \mathbb{R}^n \rightarrow \mathbb{R}^2, \quad \mathbf{x} \mapsto \begin{bmatrix} \|\mathbf{x}\|^2 - 1 \\ \mathbf{x}^\top \mathbb{1}_n \end{bmatrix}. \quad (29)$$

Then  $\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp = F^{-1}(0)$  and the derivative of  $F$  is given by

$$DF(\mathbf{x})\mathbf{h} = \begin{bmatrix} 2\mathbf{x}^\top \\ \mathbb{1}_n^\top \end{bmatrix} \mathbf{h}, \quad (30)$$

which is surjective for all  $\mathbf{x} \in \mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp$ . The regular value theorem now implies that  $\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp$  is a submanifold of  $\mathbb{R}^n$  and that  $T_{\mathbf{x}}(\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp)$  is given by the null space of  $DF(\mathbf{x})$ , yielding equation (28).  $\square$

With respect to the standard inner product, the orthogonal projection onto  $T_{\mathbf{x}}(\mathbb{S}_{n-1} \cap \mathbb{1}_n^\perp)$  is given by the projection matrix

$$\mathbf{P}_{\mathbf{x}} = (\mathbf{I}_n - \mathbf{Q}_{\mathbf{x}}\mathbf{Q}_{\mathbf{x}}^\top), \quad (31)$$

where

$$\mathbf{Q}_{\mathbf{x}} = \left[ \mathbf{x}, \frac{1}{\sqrt{n}} \mathbb{1}_n \right] \quad (32)$$

has orthonormal columns. Using the product manifold structure, we find the orthogonal projection from  $\mathbb{R}^{k \times n}$  onto  $T_{\mathbf{X}}\mathcal{R}$  as

$$\Pi_{\mathbf{X}}[\mathbf{y}_1, \dots, \mathbf{y}_k] = [\mathbf{P}_{\mathbf{x}_1}\mathbf{y}_1, \dots, \mathbf{P}_{\mathbf{x}_k}\mathbf{y}_k]. \quad (33)$$

In order to compute the Riemannian gradient of the learning function (15), note that the gradient with respect to the standard inner product is given by

$$\nabla L(\mathbf{X}_U^\top, \mathbf{X}_V^\top) = [\nabla_U L(\mathbf{X}_U^\top, \mathbf{X}_V^\top)^\top, \nabla_V L(\mathbf{X}_U^\top, \mathbf{X}_V^\top)^\top], \quad (34)$$

where  $\nabla_U$  and  $\nabla_V$  denote the gradient of  $L$  with respect to its first and second input. Using equation (33), the Riemannian gradient is thus

$$\mathbf{G}(\mathbf{X}_U, \mathbf{X}_V) = [\Pi_{\mathbf{X}_U} \nabla_U L(\mathbf{X}_U^\top, \mathbf{X}_V^\top)^\top, \Pi_{\mathbf{X}_V} \nabla_V L(\mathbf{X}_U^\top, \mathbf{X}_V^\top)^\top]. \quad (35)$$

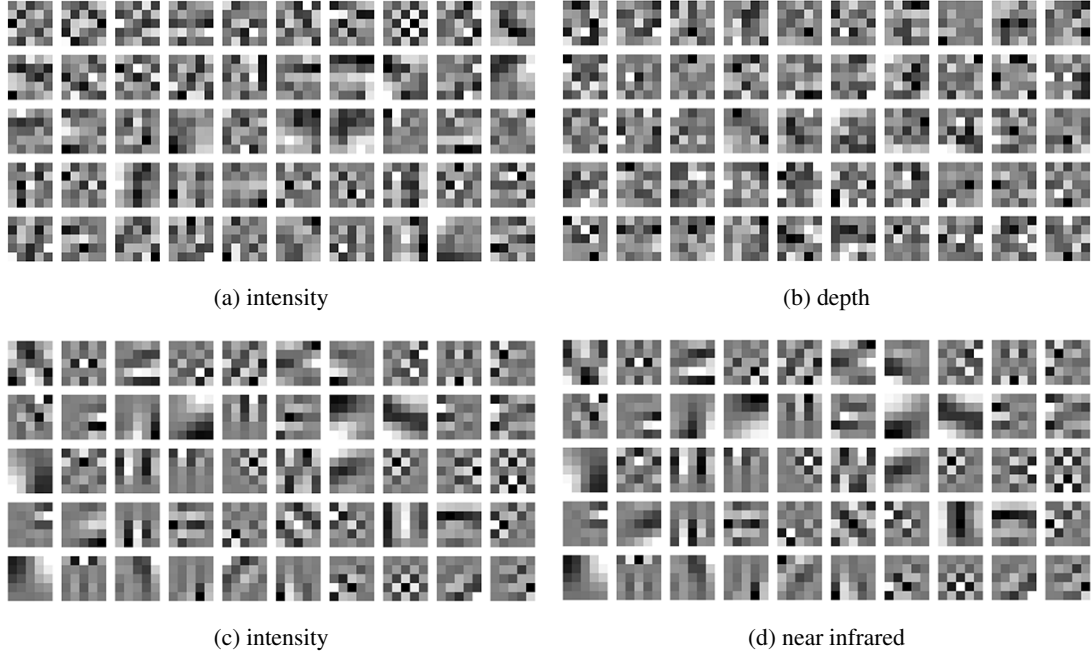


Figure 1: Plot of the rows of bimodal operator pairs visualized as square patches for intensity and depth (top row) as well as intensity and near infrared (bottom row). Patches at the same positions within their plots correspond to rows in the operators with the same index.

Since  $\mathcal{R}$  is a submanifold of the oblique manifold  $\text{OB}(n, k)$ , the formulas for the geodesics and the parallel transport coincide. We refer to [13] for explicit formulas for the geodesics and the parallel transport. Following the general conjugate gradient scheme presented in subsection 3.1, it is now straightforward to implement the learning algorithm.

Concerning the choice of training samples, we randomly select  $M$  pairs of aligned patches  $(\mathbf{s}_U^{(i)}, \mathbf{s}_V^{(i)})$  from noise-free images. Since typically, the two modalities are measured in different physical units, the patches are normalized by their standard deviation to allow a comparison of both modalities. Patches with small standard deviation are discarded for the learning process. Note, due to the restriction of the admissible set of operators to  $\mathcal{R}$ , cf. (11), patches with small standard deviation (i.e. nearly constant patches) generically fit our model and thus discarding them does not bias the learning process.

Figure 1 illustrates the rows of learned operator pairs as square patches for the two bimodal image setups intensity and depth as well as intensity and near infrared (NIR). These are used in the experiments in Section 4.3 and Section 5.2. Note how the intensity operators differ between the two setups due to the bimodal coupling with depth and infrared modalities respectively.

Learning such pairs of operators from several thousand samples on a standard desktop PC can be accomplished within the order of a few minutes.



## 4 Bimodal Image Reconstruction

The proposed model is based on learning from local patches. Yet, we want to apply this model to entire images of much larger dimension. Therefore, we first need a global formulation of the local operators to process images. Such a formulation is developed in Section 4.1. In Section 4.2, we will then show how image reconstruction can be achieved using these global operators and demonstrate its practical applicability in a super-resolution experiment in Section 4.3.

### 4.1 Application of the Patch-Based Operators to Images

According to [13], a *global* analysis operator  $\Omega^F \in \mathbb{R}^{K \times N}$  is constructed from a patch based operator  $\Omega \in \mathbb{R}^{k \times n}$  as follows. Denote the operator that extracts the normalized  $(\sqrt{n} \times \sqrt{n})$ -dimensional patch centered at position  $(r, c)$  from the entire image as  $\mathcal{P}_{rc} \in \mathbb{R}^{n \times N}$ . The global analysis operator is then given as

$$\Omega^F := \begin{bmatrix} \Omega \mathcal{P}_{11} \\ \Omega \mathcal{P}_{21} \\ \vdots \\ \Omega \mathcal{P}_{hw} \end{bmatrix} \in \mathbb{R}^{K \times N}, \quad (36)$$

with  $K = Nk$ , i.e. all patch positions are considered. We use the reflective boundary condition to deal with areas along image borders.

### 4.2 Formulation of the Bimodal Image Reconstruction Problem

The general goal of the bimodal image reconstruction task is to recover an aligned pair of bimodal images  $\mathbf{s}_U, \mathbf{s}_V \in \mathbb{R}^N$  from a set of measurements  $\mathbf{y}_U \in \mathbb{R}^{m_U}, \mathbf{y}_V \in \mathbb{R}^{m_V}$ . Here,  $\mathbf{s}_U, \mathbf{s}_V$  are the vectorized versions of the original images from each of the two modalities, obtained by ordering their entries lexicographically. Furthermore,  $w, h$  denote the height and width of both original images and  $N=wh$ .

In our reconstruction approach, we treat the problem of bimodal image reconstruction as a linear inverse problem. Formally, the relation between  $\mathbf{s}_U, \mathbf{s}_V$  and  $\mathbf{y}_U, \mathbf{y}_V$  is given by

$$\mathbf{y}_U = \Phi_U \mathbf{s}_U + \mathbf{e}_U, \quad \mathbf{y}_V = \Phi_V \mathbf{s}_V + \mathbf{e}_V. \quad (37)$$

$\Phi_U \in \mathbb{R}^{m_U \times N}, \Phi_V \in \mathbb{R}^{m_V \times N}$  model the sampling process of the measurements and  $\mathbf{e}_U \in \mathbb{R}^{m_U}, \mathbf{e}_V \in \mathbb{R}^{m_V}$  model noise and potential sampling errors. For typical reconstruction tasks, the dimensions  $m_U, m_V$  of the measurement vectors may be significantly smaller than the dimension  $N$ . Consequently, reconstructing  $\mathbf{s}_U, \mathbf{s}_V$  in (37) is highly ill-posed.

To resolve this, the bimodal data model is employed as a co-sparsity prior to regularize the image reconstruction. Accordingly, we aim to solve

$$\begin{aligned} (\mathbf{s}_U^*, \mathbf{s}_V^*) &\in \arg \min_{\mathbf{s}_U, \mathbf{s}_V \in \mathbb{R}^N} g(\Omega_U^F \mathbf{s}_U, \Omega_V^F \mathbf{s}_V) \\ \text{subject to } &d_E((\Phi_U \mathbf{s}_U, \Phi_V \mathbf{s}_V), (\mathbf{y}_U, \mathbf{y}_V)) \leq \varepsilon. \end{aligned} \quad (38)$$

We denote  $d_E$  as an appropriate data fidelity measure such as the squared Euclidean distance and  $\varepsilon \in \mathbb{R}_0^+$  is an estimated upper bound of the noise energy. Consequently, the analyzed versions of both modalities are enforced to have a correlated co-support and as a result, the two signals are coupled.

Depending on the choice of the measurement operators  $\Phi_U, \Phi_V$ , different reconstruction tasks such as denoising, inpainting, or upsampling can be performed. This can be accomplished jointly on both signals simultaneously or only on one single modality, while the other reinforces the co-sparsity and data priors. We show in the following section how image guided depth map SR can be accomplished using this model.

### 4.3 Image-guided Depth Map Super-Resolution

In this experiment, we apply the proposed reconstruction approach to the image modalities intensity and depth. Due to the availability of affordable sensors, this has become a common bimodal image setup. We now focus on recovering the HR depth image  $\mathbf{s}_D$  from LR depth measurements  $\mathbf{y}_D$ , given a fixed high quality intensity image  $\mathbf{s}_I = \mathbf{y}_I$ . In this case,  $\Phi_I$  is the identity operator and the analyzed intensity image is constant, i.e.

$$\Omega_I^F \mathbf{s}_I = \mathbf{c} = \text{const.} \quad (39)$$

This simplifies problem (38) for recovering the HR depth map to

$$\begin{aligned} \mathbf{s}_D^* \in \arg \min_{\mathbf{s}_D \in \mathbb{R}^N} g(\mathbf{c}, \Omega_D^F \mathbf{s}_D) \\ \text{subject to } d_E(\Phi_D \mathbf{s}_D, \mathbf{y}_D) \leq \varepsilon_D. \end{aligned} \quad (40)$$

The data fidelity term  $d_E$  depends on the error model of the depth data and can be chosen accordingly. For instance, this may be an error measure tailored to a sensor specific error model, as described for the Kinect sensor in [16]. In this way, knowledge about the scene gained from the intensity image and its co-support regarding the bimodal analysis operators helps to determine the HR depth signal.

To compare our results to state-of-the-art methods, we quantitatively evaluate our algorithm on the four standard test images 'Tsukuba', 'Venus', 'Teddy', and 'Cones' from the Middlebury dataset [29]. To artificially create LR input depth maps, we scale the ground truth depth maps down by a factor of  $d$  in both vertical and horizontal dimension. We first blur the available HR image with a Gaussian kernel of size  $(2d - 1) \times (2d - 1)$  and standard deviation  $\sigma = d/3$  before downsampling. The LR depth map and the corresponding HR intensity image are the input to our algorithm.

In this reconstruction from LR measurements, we assume an i.i.d. normal distribution of the error, which leads to the data fidelity term

$$d_E(\Phi_D \mathbf{s}_D, \mathbf{y}_D) = \|\Phi_D \mathbf{s}_D - \mathbf{y}_D\|_2^2. \quad (41)$$

Finally, we obtain the unconstrained formulation of problem (40) for reconstructing the HR depth image, namely

$$\mathbf{s}_D^* \in \arg \min_{\mathbf{s}_D \in \mathbb{R}^N} \lambda g(\mathbf{c}, \Omega_D^F \mathbf{s}_D) + \|\Phi_D \mathbf{s}_D - \mathbf{y}_D\|_2^2. \quad (42)$$

We solve problem (42) using a standard conjugate gradient method and an Armijo step size selection. In that matter, larger values of the weighting factor  $\lambda$  lead to a faster convergence of the algorithm but allow larger differences between the measurements and the reconstruction estimate. To achieve the best results within few iterations, we start with a large value of  $\lambda$  and restart the conjugate gradient optimization procedure several times, while consecutively shrinking the multiplier to a final value of  $\lambda = 1$ . Problem (42) is not convex and convergence to a global minimum can not be guaranteed. In practice however, we observe convergence to accurate depth maps from random initializations of  $\mathbf{s}_D$ .

For the evaluation of our approach, we train one fixed operator pair and use it in all presented intensity and depth experiments. To that end, we gather a total of  $M = 15000$  pairs of square sample patches of size  $\sqrt{n} = 5$  from the five registered intensity and depth image pairs 'Baby1', 'Bowling1', 'Moebius',

$d$	method	Tsukuba	Venus	Teddy	Cones
2x	nearest-neighbor	1.24	0.37	4.97	2.51
	Yang <i>et al.</i> [35]	1.16	0.25	2.43	<u>2.39</u>
	Diebel <i>et al.</i> [9]	2.51	0.57	2.78	3.55
	Hawe <i>et al.</i> [13] <sup>1</sup>	1.03	0.22	2.95	3.56
	JID [16]	<b>0.47</b>	<b>0.09</b>	<b>1.41</b>	<b>1.81</b>
	our method	<u>0.83</u>	<u>0.12</u>	<u>1.96</u>	2.69
4x	nearest-neighbor	3.53	0.81	6.71	5.44
	Yang <i>et al.</i>	2.56	0.42	5.95	<u>4.76</u>
	Diebel <i>et al.</i>	5.12	1.24	8.33	7.52
	Hawe <i>et al.</i>	2.95	0.65	4.80	6.54
	JID	<u>1.73</u>	<u>0.25</u>	<b>3.54</b>	5.16
	our method	<b>1.48</b>	<b>0.23</b>	<u>3.99</u>	<b>4.69</b>
8x	nearest-neighbor	3.56	1.90	10.9	10.4
	Yang <i>et al.</i>	6.95	1.19	11.50	11.00
	Diebel <i>et al.</i>	9.68	2.69	14.5	14.4
	Lu <i>et al.</i> [21]	5.09	1.00	9.87	11.30
	Hawe <i>et al.</i>	5.59	1.24	11.40	12.30
	JID	<u>3.53</u>	<b>0.33</b>	<b>6.49</b>	<u>9.22</u>
	our method	<b>3.30</b>	<u>0.34</u>	<u>8.11</u>	<b>8.57</b>

Table 1: Numerical comparison of our method to other depth map SR approaches for different upscaling factors  $d$ . The figures represent the percentage of bad pixels with respect to all pixels of the ground truth data and an error threshold of  $\delta = 1$ . Bold and underlined figures highlight the best and second best results.

$d$	method	Tsukuba	Venus	Teddy	Cones
2x	nearest-neighbor	0.612	0.288	1.543	1.531
	Chan <i>et al.</i> [5]	n/a	0.216	1.023	1.353
	Hawe <i>et al.</i> [13]eq. (40)	0.278	0.105	0.996	0.939
	JID [16]	<b>0.255</b>	<b>0.075</b>	<b>0.702</b>	<b>0.680</b>
	our method	<u>0.256</u>	<u>0.077</u>	<u>0.803</u>	<u>0.821</u>
4x	nearest-neighbor	1.189	0.408	1.943	2.470
	Chan <i>et al.</i>	n/a	0.273	<b>1.125</b>	1.450
	Hawe <i>et al.</i>	<u>0.450</u>	0.179	1.389	1.398
	JID	0.487	<u>0.129</u>	1.347	<u>1.383</u>
	our method	<b>0.374</b>	<b>0.108</b>	1.256	<b>1.287</b>
8x	nearest-neighbor	1.135	0.546	2.614	3.260
	Chan <i>et al.</i>	n/a	0.369	<b>1.410</b>	<b>1.635</b>
	Hawe <i>et al.</i>	<u>0.713</u>	0.249	1.743	1.883
	JID	0.753	<u>0.156</u>	1.662	<u>1.871</u>
	our method	<b>0.660</b>	<b>0.155</b>	<u>1.729</u>	1.931

Table 2: Numerical comparison of our method to other depth map SR approaches. The figures represent the RMSE in comparison with the ground truth depth map. Bold and underlined figures highlight the best and second best results.

‘Reindeer’ and ‘Sawtooth’ of the Middlebury stereo set [29]. Furthermore, we learn the operators with twofold redundancy, i.e.  $k = 2n$ , resulting in the operator pair  $(\Omega_I, \Omega_D) \in \mathbb{R}^{50 \times 25} \times \mathbb{R}^{50 \times 25}$ . In general, a larger redundancy of the operators leads to better reconstruction quality but at the cost of an increased

computational burden of both learning and reconstruction. Twofold redundancy provides a good trade-off between reconstruction quality and computation time. We empirically set the remaining learning parameters to  $\nu = 400$ ,  $\kappa_I = 5$ ,  $\kappa_D = 22$ ,  $\mu_I = 10^2$  and  $\mu_D = 2.5 \cdot 10^4$ .

Following the methodology described in the work of comparable depth map SR approaches, we use the Middlebury stereo matching online evaluation tool<sup>2</sup> to quantitatively assess the accuracy of our results with respect to the ground truth data. We report the percentage of bad pixels over all pixels in the depth map with an error threshold of  $\delta = 1$ . Additionally, we provide the root-mean-square error (RMSE) based on 8-bit images. We compare our results to several of the state-of-the-art methods for image guided depth map SR. Here, we focus on methods that conduct the same experiments and point the reader to [16] for a more comprehensive review of related methods.

In [35], Yang *et al.* apply bilateral filtering to depth cost volumes in order to iteratively refine an estimate using an additional color image. Chan *et al.* [5] elaborate on this approach with a fast and noise-aware joint bilateral filter. In the work of [9], color image information is used to guide depth reconstruction by computing the smoothness term in Markov-Random-Field formulation according to texture derivatives, which is extended in [21] by a data term better adapted to depth images. We also compare our results to a unimodal co-sparse analysis operator proposed by some of the authors in [13], which we learn from depth samples only. Since the unimodal approach has to solve a harder problem, this demonstrates how a bimodal approach can practically contribute to improvements in reconstruction quality. For completeness, we also include the results achieved with the joint intensity and depth (JID) method proposed earlier [16]. Where an implementation is not publicly available, we rely on the results reported by the respective authors regarding the numerical comparison in Table 1 and Table 2.

Our method improves depth map SR considerably over simple interpolation approaches. Neither stair-casing nor substantial blurring artifacts occur, particularly in areas with discontinuities. Also, there is no noticeable texture cross-talk in areas of smooth depth and cluttered intensity. Edges can be preserved with great detail due to the additional knowledge provided by the intensity image, even if SR is conducted using large upscaling factors. The quantitative comparison with other depth map SR methods demonstrates the excellent performance of our approach. The improvement over other methods is of particular significance for larger magnification factors.

## 5 Bimodal Image Registration

Image registration is the process of geometrically aligning two images that were taken by e.g. different sensors, at different points in time or from different viewpoints. Automatic image registration can be categorized into feature-based and area-based algorithms. The first group of algorithms searches for salient features in both images (e.g. edges, corners, contours) and tries to find the matching pairs of features. The geometric transformation that minimizes the distance between matching features is then used to transform one of the images. Area-based algorithms do not consider at specific features but use the whole overlapping region between both images to evaluate the registration. In both cases a distance metric is needed to either match the features or to measure the similarity between image regions. In the unimodal registration case simple metrics like the sum of squared differences or correlation can be used. Multimodal registration is more challenging because the intensities of two different sensors can differ substantially when imaging the same physical object. This phenomenon is often called contrast reversal, as bright objects in one modality can be very dark in the other and vice versa. In general, no straight-forward functional relationship between

<sup>1</sup>only takes depth as input and therefore solves a harder problem

<sup>2</sup><http://vision.middlebury.edu/stereo/eval/>

the intensities of the sensors exists. Nevertheless, the approach of Orchard [25] tries to find a piecewise linear mapping between the intensities of different modalities. The most popular metric for multimodal image registration is Mutual Information, originally introduced by Viola and Wells [31] and Collignon *et al.* [7]. A normalized version was later proposed by Studholme *et al.* [30] that is better suited to changing sizes of the overlapping region. Mutual Information is used for a variety of different applications and sensors as in medical registration [27], remote sensing [6, 11] and surveillance [18]. For more information about image registration, we refer to Brown [3] and Zitová and Flusser [36] who have published excellent overviews covering several decades of research in this area.

## 5.1 Bimodal Image Registration Algorithm

In this section we present an area-based approach that employs the formerly learned bimodal co-sparse analysis model for registration of two image modalities. We consider two images  $I_U$  and  $I_V$  of a 3D scene that are sensed through two modalities  $U$  and  $V$ . We further assume that these images can be aligned with a transformation  $\tau$  that belongs to one of the following Lie groups  $\mathcal{G}$ .

- The special orthogonal group  $SO(2)$ ;
- the special Euclidean group  $SE(2)$ ;
- the special affine group  $SA(2)$ ;
- or, the affine group  $A(2)$ .

This means that, if  $\mathbf{x}$  denotes the homogeneous pixel coordinates for one modality, say  $I_U$ , there exists some  $\tau \in \mathcal{G}$  such that the two images are perfectly aligned

$$I_V(\tau\mathbf{x}) \sim I_U(\mathbf{x}) \quad \text{for all pixel coordinates } \mathbf{x}. \quad (43)$$

Here, we have chosen the standard representation of the above groups in the set of  $(3 \times 3)$  real matrices, and the standard group action  $\tau\mathbf{x}$  on the homogeneous coordinates is simply given by a matrix-vector multiplication. Note, that the inclusions  $SO(2) \subset SE(2) \subset SA(2) \subset A(2)$  hold. We use the shorthand notation  $\tau \circ I$  for the transformed image, i.e.

$$(\tau \circ I) := I(\tau\mathbf{x}) \quad \text{for all pixel coordinates } \mathbf{x}. \quad (44)$$

The aim of this section is to find  $\tau$  by using the bimodal pair of analysis operators  $(\Omega_U, \Omega_V)$ . The idea behind our approach is, that for an optimal transformation, the coupled sparsity measure should be minimized. Thus, we are searching for  $\tau^* \in \mathcal{G}$  such that

$$\tau^* \in \arg \min_{\tau \in \mathcal{G}} g\left(\Omega_U^F I_U, \Omega_V^F (\tau \circ I_V)\right) \quad (45)$$

In order to tackle the above optimization problem, we follow an approach that is similar to what has been proposed in [26]. It is based on iteratively updating the estimate of  $\tau$  with group elements near the identity. Locally, the Matrix exponential yields a diffeomorphism between a neighborhood of the identity in  $\mathcal{G}$  and a neighborhood around 0 in the corresponding Lie algebra  $\mathfrak{g}$  of  $\mathcal{G}$ . For the considered Lie groups at hand, each Lie algebra is contained in

$$\mathfrak{g} := \left\{ \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ 0 & 0 \end{bmatrix} \mid \mathbf{A} \in \mathbb{R}^{2 \times 2}, \mathbf{b} \in \mathbb{R}^2 \right\}, \quad (46)$$

which is the Lie algebra of  $A(2)$ . Further restrictions on the parameters then lead to the corresponding Lie algebras for the sub groups. For  $SO(2)$ , we have  $\mathbf{A}^\top = -\mathbf{A}$  and  $\mathbf{b} = 0$ , for  $SE(2)$  we have  $\mathbf{A}^\top = -\mathbf{A}$ , and for  $SA(2)$  we have  $\text{tr } \mathbf{A} = 0$ .

Thus, for a transformation  $\delta$  which is near the identity, we have  $\delta = e^{\mathbf{H}}$  for some matrix  $\mathbf{H} \in \mathfrak{g}$  in a neighborhood of 0. Now, in order to tackle the optimization problem (45) we proceed as follows. For legibility, denote

$$F(\tau) := g\left(\Omega_U^F I_U, \Omega_V^F(\tau \circ I_V)\right). \quad (47)$$

We employ a geometric gradient descent method described in Section 3.1 on the Lie group  $\mathcal{G}$  for minimizing  $F(\tau)$  that updates  $\tau$  in each step. To that end, we endow the set of  $(3 \times 3)$  real matrices with the inner product

$$\langle \mathbf{H}_1, \mathbf{H}_2 \rangle_{\mathbf{P}} := \text{tr} \left( (\mathbf{H}_1 \odot \mathbf{P}) \mathbf{H}_2^\top \right), \quad (48)$$

with  $\mathbf{P}$  having positive entries and  $\odot$  denoting the Hadamard product. The choice of  $\mathbf{P}$  allows to balance the translational versus the rotational part of the chosen group, or the shearing part, respectively. This is commonly done to account for different magnitudes of the transformation parameters [17].

Choose  $\tau_0 := \text{id}$  as an initialization. Then iterate the following steps until convergence.

1. Compute the Riemannian Gradient of  $F(\delta \circ \tau)$  at  $\delta = \text{id}$ , which is an element of the Lie algebra

$$\mathbf{G} := \text{grad}_\delta F(\delta \circ \tau) \big|_{\delta=\text{id}} \in \mathfrak{g}. \quad (49)$$

2. Choose an approximate step-size  $t^*$  for

$$\phi(t) = F(e^{t\mathbf{G}} \circ \tau). \quad (50)$$

3. Update  $\tau \leftarrow e^{t^*\mathbf{G}} \tau$ .

For our problem at hand, we choose the Armijo rule to determine the step size. We refer to the Appendix for the derivation of the gradient of  $F(\delta \circ \tau)$ . As a stopping criterion, we choose a threshold for the norm of the Riemannian gradient.

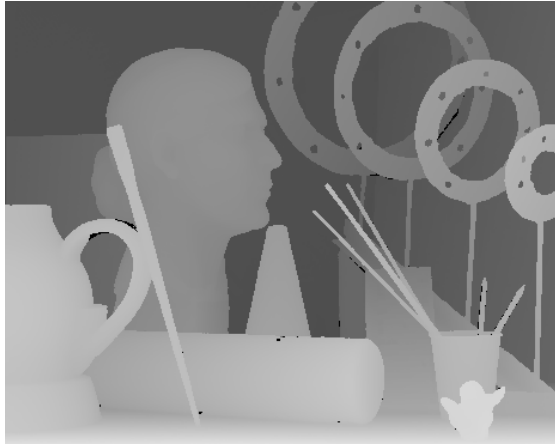
## 5.2 Evaluation

We compare our registration approach to two multimodal registration metrics, namely Mutual Information (MI) and Normalized Mutual Information (NMI) [23]. The elastix image registration toolbox [17] provides the reference implementations of these metrics together with a gradient descent algorithm to find the transformation parameters. In all cases we use the standard parameters of the elastix toolbox. In our experiments, we use intensity and depth images from the Middlebury stereo set and images from the RGB-NIR Scene Dataset [4]. The RGB-NIR dataset consists of RGB images and near-infrared (NIR) images that were captured with commercial DSLR cameras using filters for the visible and infrared spectrum. The spectra do not overlap (the cutoff wavelength is about 750 nm) and the NIR images give statistically different information from the R, G and B channel. Both datasets are very well registered and we use this registration as the ground truth and learn the operators on registered training images.

We train one fixed operator pair for each of the registration scenarios intensity+depth and intensity+NIR. For the intensity and depth setup, we use the same operator as in the reconstruction experiments in Section



(a) art intensity image



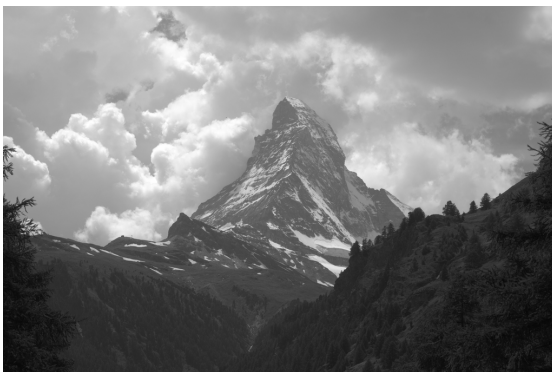
(b) art depth map



(c) intensity old building



(d) NIR old building



(e) intensity mountain



(f) NIR mountain

Figure 2: Example images used in the registration experiments. The intensity and depth image pair differ significantly and is challenging for multimodal registration algorithms. The NIR images are more similar to the intensity images and differ mainly in areas with vegetation and sky.

deregistration (x,y, $\theta$ )	method	intensity-depth (art)	intensity-NIR (old building)	intensity-NIR (mountain)
0, 0, 10	MI	14.70, 50.14, -2.01	6.14, 1.56, -3.01	<b>0.16, -0.85, -3.52</b>
	NMI	9.42, 18.66, -7.83	2.49, 4.26, -9.17	-1.43, -1.05, -3.85
	our method	<b>-1.11, -1.41, 0.11</b>	<b>0.35, 0.22, 0.01</b>	-0.34, -2.95, -8.31
-5, -5, 0	MI	-1.06, <b>1.13</b> , 0.02	-0.21, -0.14, 0.05	0.10, -0.06, <b>0.05</b>
	NMI	7.02, 4.96, <b>0.01</b>	<b>0.03, 0.10, 0.02</b>	<b>0.05, 0.01</b> , 0.06
	our method	<b>-1.00</b> , -1.13, 0.03	-1.03, 2.34, 2.72	2.56, 0.41, 0.06
10, 0, 5	MI	8.60, 18.71, -2.49	-8.59, 2.26, -1.32	2.64, 0.08, -1.76
	NMI	3.44, 9.79, -3.27	-8.22, 2.27, -1.35	2.58, <b>0.05</b> , -1.71
	our method	<b>-0.90, -0.81, 0.19</b>	<b>0.02, 0.23, 0.06</b>	<b>0.61</b> , 0.37, <b>-0.02</b>
-5, -5, 5	MI	1.38, 11.12, -2.65	-7.53, 1.79, -1.43	1.57, -0.23, -1.77
	NMI	3.04, 8.87, -3.01	-8.83, 1.86, -1.36	1.58, -0.27, -1.73
	our method	<b>-0.22, -1.40, 0.28</b>	<b>0.22, -0.13, 0.14</b>	<b>0.82, -0.01, -0.28</b>

Table 3: Registration residual for different synthetic translations and rotations. Values for the translation in x and y direction are given in pixels, the angle  $\theta$  is given in degrees.

4.3. For the intensity and NIR setup, we followed the same learning procedure, randomly collecting  $M = 15000$  pairs of square sample patches of size  $\sqrt{n} = 5$  from a total of 9 images in the training set, one from each category which we then exclude from testing. We set the learning parameters to  $\nu = 200$ ,  $\kappa_I = 250$ ,  $\kappa_N = 1000$ ,  $\mu_I = 250$  and  $\mu_N = 1000$ . All other parameters are the same as for intensity and depth.

In order to evaluate the result of the registration of one image pair, we apply a synthetic deregistration to one of the images. This deregistration consists of a translation and a rotation and subsequently the registration algorithm searches for a transformation that belongs to the special Euclidian group. Both the elastix toolbox and our algorithm work on a Gaussian image pyramid of four levels.

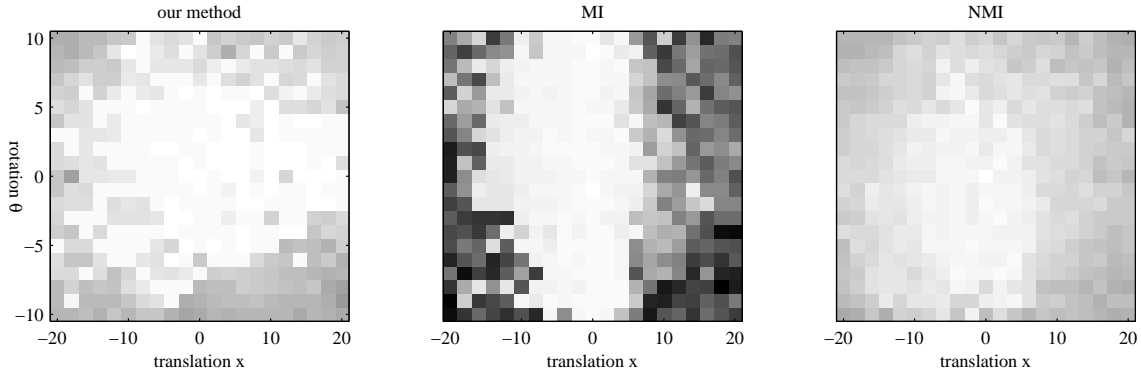


Figure 3: Remaining combined registration error for different initial deregistrations which consist of a translation in x-direction and a rotation  $\theta$ . White and black areas correspond to small and large errors, respectively. MI fails to register the images for large translations. Our method achieves the smallest remaining error and can handle large translations and rotations.

Table 3 shows the remaining registration error after running the different registration algorithms. Our method achieves comparable or better results than MI or NMI for all of the modalities. The MI and NMI



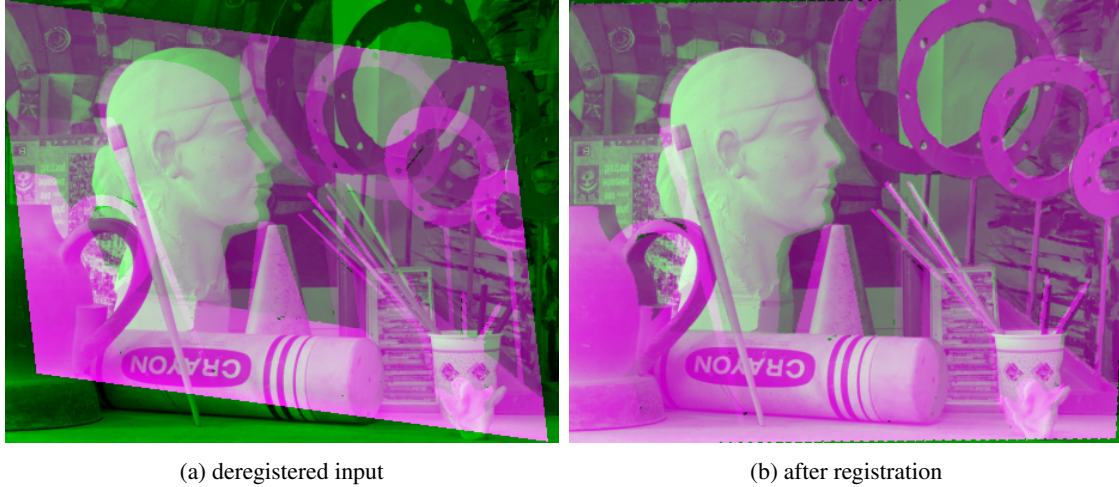


Figure 4: Example of an intensity and depth image pair before and after the registration process using an affine transformation.

algorithms fail to register the intensity and depth image pair in most of the cases but achieve better results in the intensity-NIR images. This can be explained by the fact that intensity and depth are much less alike than intensity and near-infrared (see Figure 2). Figure 3 shows the remaining registration error for various initial deregistrations of the intensity and depth image pair. We define the remaining combined registration error as

$$\epsilon = \sqrt{\epsilon_x^2 + \epsilon_y^2 + \epsilon_\theta^2}, \quad (51)$$

where  $\epsilon_x$  and  $\epsilon_y$  denote the remaining translation error in x- and y-direction (in pixels) and  $\epsilon_\theta$  denotes the remaining rotation error (in degrees). White areas in Figure 3 correspond to small registration errors ( $\epsilon < 1$ ) and black areas show large errors ( $\epsilon > 50$ ) where the registration has failed. It can be seen that MI is susceptible to large translations and fails to align the images correctly. The direct comparison of our method and NMI shows that both algorithms can handle the initial deregistration better than MI but our method achieves smaller remaining errors over a wider range of deregistration values.

## 6 Conclusion and Discussion

We have introduced a way to model the interdependencies of two image modalities by extending the co-sparse analysis model. The coupled analysis operators are learned by minimizing a coupled sparsity function via a conjugate gradient method on an appropriate manifold. The manifold setting allows to constrain the rows of the operators a priori to zero mean and unit norm, which accounts for the intuition that contrast in image modalities is most informative.

We evaluate the descriptive power of the presented model in two application scenarios. First, we have used it as a regularizer for inverse problems in imaging and provided numerical experiments for depth map super-resolution, when a high resolution intensity image is available for the same scene. As a second application scenario, we have considered the problem of bimodal image registration. An algorithm on Lie

groups has been proposed that uses a previously learned bimodal model to register intensity and depth images as well as intensity and near infrared (NIR) images.

Despite these convincing results, using our model for bimodal image processing certainly has some limitations. The model is based on local assumptions and thus fails to regularize tasks where large areas of one modality are not available. For example, inpainting of large gaps in the image typically fails due to the lack of global support, as is also reported in [16].

Now clearly, the introduction of a new model poses at least as many questions as it does provide answers. Our evaluation shows promising results for intensity/depth and intensity/NIR images. The question how our model will perform for other modality compositions remains open at this point. For example, it would be interesting to investigate its applicability in medical imaging, where different modalities like MRI or PET play an important role. We leave these questions to future work.

## A Appendix

### A.1 Derivation of the Riemannian gradient in Section 5

In this section we derive the Riemannian gradient in Eq. (49) for the bimodal alignment algorithm. We make use of the following criterion for its derivation. Let  $\langle \cdot, \cdot \rangle_{\mathbf{P}}$  be the Riemannian metric on the Lie group  $\mathcal{G}$  inherited from (48) and let  $F(\cdot)$  be a smooth real valued function on  $\mathcal{G}$ . Then the Riemannian gradient of  $F$  at  $\delta \in \mathcal{G}$  is the unique vector  $\mathbf{G} \in T_{\delta}\mathcal{G}$ , the tangent space at  $\delta$ , such that

$$\left. \frac{d}{dt} \right|_{t=0} F(e^{t\mathbf{H}}\delta) = \langle \mathbf{H}, \mathbf{G} \rangle_{\mathbf{P}} \quad (52)$$

holds for all tangent elements  $\mathbf{H} \in T_{\delta}\mathcal{G}$ .

For our purpose, we compute the gradient at  $\delta = \text{id}$ . Now let  $B$  be the image region in which we want to align the modalities  $I_U$  and  $I_V$ . We assume that  $B$  is rectangular and denote by

$$I(\mathbf{x})_{\mathbf{x} \in B} \quad (53)$$

the vectorized version of  $I$  over the domain  $B$ . Using Equation (47) and the fact that  $\mathbf{c} := \Omega_U^F I_U$  is a constant vector, we compute by the chain rule that

$$\begin{aligned} \left. \frac{d}{dt} \right|_{t=0} F(e^{t\mathbf{H}}\tau) &= \left. \frac{d}{dt} \right|_{t=0} g(\mathbf{c}, \Omega_V^F [(e^{t\mathbf{H}}\tau) \circ I_V]) \\ &= \nabla g(\mathbf{c}, \Omega_V^F I_V(\tau\mathbf{x})_{\mathbf{x} \in B})^\top \Omega_V^F \left[ \left. \frac{d}{dt} \right|_{t=0} I_V(e^{t\mathbf{H}}\tau\mathbf{x})_{\mathbf{x} \in B} \right]. \end{aligned} \quad (54)$$

The last bracket is a vector where each of its entries is computed as

$$\left. \frac{d}{dt} \right|_{t=0} I_V(e^{t\mathbf{H}}\tau\mathbf{x}) = \nabla I_V(\tau\mathbf{x})^\top \mathbf{H}\tau\mathbf{x} = \text{vec}(\tau\mathbf{x} \otimes \nabla I_V(\tau\mathbf{x}))^\top \text{vec}(\mathbf{H}), \quad (55)$$

where as usual,  $\text{vec}(\cdot)$  denotes the linear operator that stacks the columns of a matrix among each other and  $\otimes$  is the Kronecker product. Note, that since we stick to the representation with homogeneous coordinates,  $\nabla I_V(\mathbf{x}) \in \mathbb{R}^3$  is the common image gradient of  $I_V$  with an additional 0 in the third component.

Thus with

$$\mathbf{r}^\top := \nabla g(\mathbf{c}, \Omega_V^F I_V(\tau\mathbf{x})_{\mathbf{x} \in B})^\top \Omega_V^F (\text{vec}(\tau\mathbf{x} \otimes \nabla I_V(\tau\mathbf{x}))^\top)_{\mathbf{x} \in B}, \quad (56)$$

we have

$$\left. \frac{d}{dt} \right|_{t=0} F(e^{t\mathbf{H}}\delta) = \mathbf{r}^\top \text{vec}(\mathbf{H}) = \text{tr}(\text{vec}^{-1}(\mathbf{r})\mathbf{H}^\top) = \langle \text{vec}^{-1}(\mathbf{r}) \odot \hat{\mathbf{P}}, \mathbf{H} \rangle_{\mathbf{P}}, \quad (57)$$

where the entries of  $\hat{\mathbf{P}}$  are the inverse of the entries of  $\mathbf{P}$ .

Using Equation (52), the Riemannian gradient is therefore the orthogonal projection of  $\text{vec}^{-1}(\mathbf{r}) \odot \hat{\mathbf{P}}$  with respect to  $\langle \cdot, \cdot \rangle_{\mathbf{P}}$  onto the tangent space of  $\delta = \text{id}$ , which is nothing else than the Lie algebra  $\mathfrak{g}$ , i.e.

$$\text{grad}_\delta F(\delta \circ \tau) = \Pi_{\mathfrak{g}} \left( \text{vec}^{-1}(\mathbf{r}) \odot \hat{\mathbf{P}} \right). \quad (58)$$

If we further assume for the entries  $p_{ij}$  of  $\mathbf{P}$  that

$$p_{11} = p_{22} \text{ and } p_{12} = p_{21}, \quad (59)$$

then, for the considered Lie groups, these projections are explicitly given by

$$\Pi_{SO}(\mathbf{X}) = \begin{bmatrix} \frac{1}{2}(\mathbf{X}_{11} - \mathbf{X}_{11}^\top) & 0 \\ 0 & 0 \end{bmatrix} \quad (60)$$

$$\Pi_{SE}(\mathbf{X}) = \begin{bmatrix} \frac{1}{2}(\mathbf{X}_{11} - \mathbf{X}_{11}^\top) & \mathbf{x}_{12} \\ 0 & 0 \end{bmatrix} \quad (61)$$

$$\Pi_{SA}(\mathbf{X}) = \begin{bmatrix} (\mathbf{X}_{11} - \frac{1}{2} \text{tr}(\mathbf{X}_{11})\mathbf{I}_2) & \mathbf{x}_{12} \\ 0 & 0 \end{bmatrix} \quad (62)$$

$$\Pi_A(\mathbf{X}) = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{x}_{12} \\ 0 & 0 \end{bmatrix}, \quad (63)$$

where  $\mathbf{X} \in \mathbb{R}^{3 \times 3}$  is partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{x}_{12} \\ \mathbf{x}_{21}^\top & x_{22} \end{bmatrix}. \quad (64)$$

## References

- [1] Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2008)
- [2] Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(9), 1167–1183 (2002)
- [3] Brown, L.G.: A survey of image registration techniques. ACM Computing Surveys (CSUR) **24**, 325–376 (1992)
- [4] Brown, M., Süssstrunk, S.: Multi-spectral SIFT for scene category recognition. In: Computer Vision and Pattern Recognition (CVPR11), pp. 177–184 (2011)
- [5] Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A Noise-Aware Filter for Real-Time Depth Upsampling. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (2008)

- [6] Cole-Rhodes, A.A., Johnson, K.L., LeMoigne, J., Zavorin, I.: Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. *IEEE Transactions on Image Processing* **12**(12), 1495–511 (2003)
- [7] Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., Marchal, G.: Automated multi-modality image registration based on information theory. In: *Information processing in medical imaging*, vol. 3, pp. 263–274 (1995)
- [8] Dai, Y., Yuan, Y.: An efficient hybrid conjugate gradient method for unconstrained optimization. *Annals of Operations Research* **103**(1-4), 33–47 (2001)
- [9] Diebel, J., Thrun, S.: An Application of Markov Random Fields to Range Sensing. In: *NIPS*, vol. 18, pp. 291–298 (2005)
- [10] Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus Synthesis in Signal Priors. *Inverse Problems* **23**(3), 947–968 (2007)
- [11] Fan, X., Rhody, H., Saber, E.: A Spatial-Feature-Enhanced MMI Algorithm for Multimodal Airborne Image Registration. *IEEE Transactions on Geoscience and Remote Sensing* **48**(6), 2580–2589 (2010)
- [12] Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning Low-Level Vision. *International Journal of Computer Vision* **40**(1), 25–47 (2000)
- [13] Hawe, S., Kleinsteuber, M., Diepold, K.: Analysis Operator Learning and Its Application to Image Reconstruction. *IEEE Transactions on Image Processing* **22**(6), 2138–2150 (2013)
- [14] Hong Chang, Dit-Yan Yeung, Yimin Xiong: Super-resolution through neighbor embedding. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. 275–282. IEEE (2004)
- [15] Jia, K., Wang, X., Tang, X.: Image transformation based on learning dictionaries across image spaces. *IEEE transactions on pattern analysis and machine intelligence* **35**(2), 367–80 (2013)
- [16] Kiechle, M., Hawe, S., Kleinsteuber, M.: A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013* (2013)
- [17] Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging* **29**(1), 196–205 (2010)
- [18] Krotosky, S.J., Trivedi, M.M.: Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding* **106**(2-3), 270–287 (2007)
- [19] Li, Y., Xue, T., Sun, L., Liu, J.: Joint Example-Based Depth Map Super-Resolution. In: *IEEE International Conference on Multimedia and Expo*, pp. 152–157 (2012)
- [20] Liu, C., Shum, H.Y., Freeman, W.T.: Face Hallucination: Theory and Practice. *International Journal of Computer Vision* **75**(1), 115–134 (2007)
- [21] Lu, J., Min, D., Pahwa, R.S., Do, M.N.: A Revisit to MRF-Based Depth Map Super-Resolution and Enhancement. In: *ICASSP*, pp. 985–988 (2011)

- [22] Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence* **34**(4), 791–804 (2012)
- [23] Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W.: PET-CT image registration in the chest using free-form deformations. *IEEE Transactions on Medical Imaging* **22**(1), 120–128 (2003)
- [24] Nam, S., Davies, M.E., Elad, M., Gribonval, R.: The Cospase Analysis Model and Algorithms. *Applied and Computational Harmonic Analysis* **34**(1), 30–56 (2013)
- [25] Orchard, J.: Efficient least squares multimodal registration with a globally exhaustive alignment search. *IEEE Transactions on Image Processing* **16**(10), 2526–2534 (2007)
- [26] Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence* **34**(11), 2233–46 (2012)
- [27] Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging* **22**(8), 986–1004 (2003)
- [28] Ravishanker, S., Bresler, Y.: Learning Sparsifying Transforms. *IEEE Transactions on Signal Processing* **61**(5), 1072–1086 (2013)
- [29] Scharstein, D., Szeliski, R.: High-Accuracy Stereo Depth Maps using Structured Light. In: *CVPR*, pp. 195–202 (2003)
- [30] Studholme, C., Hill, D., Hawkes, D.: An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition* **32**(1), 71–86 (1999)
- [31] Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. *International Journal of Computer Vision* **24**(2), 137–154 (1997)
- [32] Wang, S., Zhang, D., Liang, Y., Pan, Q.: Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2216–2223. *IEEE* (2012)
- [33] Yaghoobi, M., Nam, S., Gribonval, R., Davies, M.E.: Constrained Overcomplete Analysis Operator Learning for Cospase Signal Modelling. *IEEE Transactions on Signal Processing* **61**(9), 2341–2355 (2013)
- [34] Yang, J., Wright, J., Huang, T., Ma, Y.: Image Super-Resolution via Sparse Representation. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **19**(11), 2861–2873 (2010)
- [35] Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-Depth Super Resolution for Range Images. In: *CVPR*, pp. 1–8 (2007)
- [36] Zitová, B., Flusser, J.: Image registration methods: a survey. *Image and Vision Computing* **21**(11), 977–1000 (2003)